EdWorkingPaper No. 25-1352

Does State-Mandated Third-Grade Reading Retention Policy Improve Achievement? Evidence from a Staggered-Adoption Difference-in-Differences Design

Sakib Mahmud
Florida State University

This paper investigates whether the state-mandated third-grade reading retention policy autonomously enhances student achievement or depends on broader literacy reforms. Using district-level data from the Stanford Education Data Archive (2010–2019), I employ a staggered-adoption Difference-in-Differences design, as per Callaway and Sant'Anna (2021), to assess heterogeneous treatment effects across four adopting states: Arizona, North Carolina, Ohio, and Mississippi. Results indicate that states implementing retention mandates observe an increase of approximately 0.07–0.10 standard deviations in fourth-grade Reasoning through Language Arts (RLA) scores. Nonetheless, when the mandate is examined in isolation, compared to states with similar literacy initiatives that allow local discretion over retention, the effect diminishes to 0.014 and becomes statistically insignificant. Mississippi is different because of its comprehensive Literacy-Based Promotion Act, which combined retention with intensive teacher training, literacy coaches, and diagnostic supports, leading to larger gains in grade 4 (0.10–0.16 SD). The results indicate that the state-mandated retention policy, by itself, does not lead to better performance; improvements occur only when mandates are part of larger, well-guided literacy programs.

VERSION: December 2025

Suggested citation: Mahmud, Sakib. (2025). Does State-Mandated Third-Grade Reading Retention Policy Improve Achievement? Evidence from a Staggered-Adoption Difference-in-Differences Design. (EdWorkingPaper: 25-1352). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/b0em-9a84

Does State-Mandated Third-Grade Reading Retention Policy Improve Achievement? Evidence from a Staggered-Adoption Difference-in-Differences Design[1]

Sakib Mahmud

Department of Economics, Florida State University, Tallahassee, Florida, Email: mahmudsakib99@gmail.com

Abstract

This paper investigates whether the state-mandated third-grade reading retention policy autonomously enhances student achievement or depends on broader literacy reforms. Using district-level data from the Stanford Education Data Archive (2010–2019), I employ a staggered-adoption Difference-in-Differences design, as per Callaway and Sant'Anna (2021), to assess heterogeneous treatment effects across four adopting states: Arizona, North Carolina, Ohio, and Mississippi. Results indicate that states implementing retention mandates observe an increase of approximately 0.07–0.10 standard deviations in fourth-grade Reasoning through Language Arts (RLA) scores. Nonetheless, when the mandate is examined in isolation, compared to states with similar literacy initiatives that allow local discretion over retention, the effect diminishes to 0.014 and becomes statistically insignificant. Mississippi is different because of its comprehensive Literacy-Based Promotion Act, which combined retention with intensive teacher training, literacy coaches, and diagnostic supports, leading to larger gains in grade 4 (0.10–0.16 SD). The results indicate that the state-mandated retention policy, by itself, does not lead to better performance; improvements occur only when mandates are part of larger, well-guided literacy programs.

JEL Classification

H52, I20, I21, I28, I29

Keywords

Early Literacy Laws; Economics of Education; Public Schools; Grade Retention

---

[1] This working paper was last updated on December 03, 2025. This paper was completed as part of the graduate program in Economics at Florida State University. I thank the faculty for valuable guidance and encouragement. All errors are my own.

1.0 Introduction

Grade retention is the practice of having students who fail to meet academic benchmarks repeat the same grade rather than be promoted. It is a longstanding, but contentious strategy used by educators and policymakers to address performance disparities in K–12 education (Goos et al., 2021). The policy is critical in the early grades, when basic reading and writing skills are crucial for future learning and social and economic success. Children who struggle to read proficiently are more likely to experience long-term academic underachievement and face adverse life outcomes, including higher risks of school dropout, unemployment, and incarceration (Fiester, 2013; Sparks et al., 2014). For example, the Arizona Department of Education reports that students who fail to read proficiently by third grade are four times more likely to drop out of high school, and 90 percent of high school dropouts struggled with reading at that stage (Read On Arizona, 2018).

As reading proficiency by the end of third grade is widely regarded as a pivotal milestone for future academic success (Cunningham & Stanovich, 1997; Hernandez, 2011), early literacy has become a significant focus of education policy over the past two decades. By 2021, 41 states and the District of Columbia had adopted early literacy laws (ExcelinEd, 2021). Despite widespread adoption, early literacy policies vary substantially across states in design and enforcement. According to ExcelinEd (2021), state laws differ in the inclusion of four major policy components: (1) teacher training and implementation support, (2) assessment and parental notification, (3) classroom instruction and intervention, and (4) retention and intensive remediation.

Roughly 10 percent of American students are retained at least once between kindergarten and eighth grade, with the practice disproportionately affecting low-income and minority students (Planty et al., 2008). Proponents argue that retention aligns students' academic readiness with classroom expectations, allowing them to master foundational skills before advancing. Critics counter that it can stigmatize students, lower teacher expectations, disrupt peer relationships, and ultimately reduce educational attainment.

Psychological and sociological research has often found negative associations between grade retention and later academic or social outcomes. Although recent economic studies using more credible identification strategies, such as Jacob and Lefgren (2009), Greene and Winters (2007), and Schwerdt et al. (2017), report smaller or even positive short-term effects of early-grade retention on student achievement. Nevertheless, it remains unclear whether state-mandated retention requirements themselves improve achievement beyond the broader instructional reforms that accompany them, a gap this study seeks to address.

This paper employs district-level data from the Stanford Education Data Archive (SEDA) to delineate the causal effects of state-mandated third-grade retention policies on student achievement in the United States. The study looks at scores in Reasoning Through Language Arts (RLA) for fourth graders from 2010 to 2019. Arizona, North Carolina, Ohio, and Mississippi are the four treatment states because each adopted a statewide third-grade retention mandate as part of a broader early-literacy reform between 2010 and 2019. Among my treatment states with state-mandated retention policy in effect, the actual retention rate varies from 1% to 14% (Loewus, 2015; Berne, Jordan S. et al., 2025; Churchill, 2023; James, 2024; Perrault & Winters, 2020). Although states with a state-mandated retention rule are stricter, good cause exemptions reduce the actual number of students retained (Berne, Jordan S. et al., 2025).

Beyond retention, literacy policies include several reading interventions, such as summer reading programs for struggling students, individualized improvement plans, and regular progress monitoring (Cummings & Turner, 2020). Among these states, Mississippi implemented the most comprehensive model through its Literacy-Based Promotion Act (LBPA), enacted in 2013. The LBPA expanded interventions to grades K–3, required teachers to get training in evidence-based reading instruction, and sent state-funded literacy coaches to schools that were not doing well.

Employing a staggered-adoption Difference-in-Differences framework as delineated by Callaway and Sant'Anna (2021), I calculate dynamic and cohort-specific Average Treatment Effects on the Treated (ATETs). Results indicate that exposure to a state retention policy enhances fourth-grade reading achievement by 0.07–0.10 standard deviations (Models 1–2). When the comparison is limited to states with robust literacy initiatives with locally governed retention (Model 3), the effect diminishes and becomes statistically insignificant (0.014). These findings indicate that improvements in achievement arise from comprehensive literacy interventions rather than solely from the retention mandate, aligning with evidence from Westall and Cummings (2023) and Spencer (2024). It is worth mentioning that the results are powerful for Mississippi, showing gains of 0.10–0.16 standard deviations across specifications and retaining a significant positive effect (0.041) even after isolating the mandate. This persistence suggests that Mississippi's success is primarily due to the intensity of its implementation: teacher coaching, professional development, and early intervention infrastructure rather than the policy's retention aspect. The contribution of this paper is to differentiate the impact of the state-mandated retention mechanism from the accompanying instructional supports, showing that improvements in literacy are driven primarily by sustained, high-quality intervention rather than by state-mandated retention enforcement itself.

The paper proceeds as follows. In Section 2, I discuss the empirical estimates from the related literature to establish the gap. In Section 3, I describe my control and treatment states, along with the data and models used for identification. Section 4 presents my three models: one with no retention as the control group, with no anticipation assumption; one with anticipation for MS but the same control group; and the last for isolating the effect of state-mandated retention. In Section 5, I compare the model's findings with previous literature and explain the mechanism underlying grade improvement. Section 6 concludes.

2.0 Related Literature

There is promising empirical literature on the impact of early literacy policy. Retention is part of early literacy policy; the most pertinent paper examining the effects of early literacy legislation throughout the United States is by Westall and Cummings (2023). The research indicates that cohorts exposed to "comprehensive" early literacy policies since kindergarten experience an approximate increase of 5 points in grade 4 NAEP reading test scores, a similar increase of 5 points in grade 4 NAEP math test scores, an enhancement of approximately 0.10 standard deviations in grade 3 SEDA RLA scores, and an improvement of roughly 0.20 standard deviations in grade 3 SEDA math scores. Spencer (2024) enhances this study by focusing on a particularly compelling example of a holistic early literacy policy in Mississippi, the Literacy-Based Promotion Act, which features teacher training and coaching as a significant distinguishing feature[2]. The author estimates that, for the average student exposed since kindergarten, the LBPA elevated grade 4 NAEP reading test scores by 8 points, enhanced grade 4 NAEP math test scores by 9 points, improved grade 3 SEDA RLA scores by 0.26 standard deviations, and increased grade 3 SEDA math scores by 0.41 standard deviations.

Meanwhile, regarding grade retention, meta-analyses and systematic reviews indicate that grade retention is less effective than educators commonly perceive, particularly in the long term. Goos et al. (2021) conducted an analysis indicating that, on average, grade retention has no significant effect globally. Nonetheless, relative to other nations, US grade retention policies are the most efficacious. The United States often handles student diversity through ability grouping, setting, and streaming (Goos et al., 2021). It is less common for students to repeat a grade, and when it does occur, it is often associated with extra help during the retention year (Schwerdt, West, & Winters, 2017). In this scenario, grade retention may yield beneficial outcomes for repeaters' academic performance through rigorous remediation, while simultaneously producing adverse effects on their psychosocial functioning, as if they were the only individuals required to repeat.

Research on grade retention in the USA primarily focuses on a limited number of individual state evaluations. Greene & Winters (2007, 2009) conducted several quasi-experimental analyses of Florida's early literacy policy, predominantly using discontinuities created by its retention component. They observe beneficial short-term impacts on students' reading performance in state assessments, along with favorable spillover effects on third-grade mathematics achievement; however, these effects diminish as students advance in their education (Greene & Winters, 2007, 2009; Schwerdt et al., 2017). Florida's fourth-grade reading scores on the NAEP improved due to its early literacy policy; however, it remains ambiguous whether retention or other policy elements contributed to these enhancements (Duke et al., 2014). Greene and Winters (2007) found that third-grade retention enhanced student achievement after 2 years, while Winters and Greene (2012) provided evidence from same-grade comparisons indicating that these improvements persisted through eighth grade. Ozek (2015) examines behavioral outcomes and discovers that students

[2] In the literature, the evidence regarding teacher training programs is inconclusive, with specific studies indicating beneficial outcomes (Angrist & Lavy, 2001; Cilliers, Fleisch, Prinsloo, & Taylor, 2020), others revealing no significant effects (Feng & Sass, 2013), and additional research demonstrating variable effects (Bassi, Meghir, & Reynoso, 2020).

retained under the Florida policy experienced increased discipline in the initial two years post-retention; however, these effects completely diminished after two years. Jacob and Lefgren (2009) examine the effects of retention in third, sixth, and eighth grades on academic achievement and high school graduation rates in Chicago. They discovered that retention in higher grade levels may adversely affect future student outcomes, whereas early grade retention may prove more advantageous. Moreover, Schwerdt et al. (2017) affirm that test-based retention in third grade in Florida enhances students' short-term academic performance; however, these initial advantages diminish over time. Simultaneously, Schwerdt et al. (2017) further provide evidence that test-based retention in third grade diminishes the likelihood of retention in subsequent grades, underscoring an additional consequence of policies that elevate retention rates in early grades. To the best of my knowledge, no paper examines the impact of state-mandated retention on achievement. After identifying the gap, my work contributes as the inaugural study examining the impact of state-mandated third-grade retention policy, in isolation from other reading interventions, on academic achievement.

3.0 Methodology

I estimate the Average Treatment Effect on the Treated (ATET) over time and across different treatment cohorts. I define a treatment cohort as a set of states that are subject to treatment at different points in time (*Table 1*). After that exposure, they remain exposed to the treatment (with no reversibility).

Table 1: Required Retention by Treatment States

State	Announced	3rd Grade First Cohort	4th Grade First Cohort
Arizona	**2010** (Move On When Reading)	2013–14	2014–15
North Carolina	**2012** (Read to Achieve)	2013–14	2014–15
Ohio	**2012** (Third Grade Reading Guarantee)	2013-14	2014-15
Mississippi	**2013** (Literacy-Based Promotion Act)	2014–15	2015–16

The treatment and control states are selected based on the report of Cummings and Turner (2020). For Models 1 and 2, control states include states that did not effectively implement the 3rd-grade retention policy up to the year 2019; those include Wisconsin, Wyoming, Utah, Vermont, Virginia, South Dakota, Rhode Island, Pennsylvania, Oregon, North Dakota, New York, New Hampshire, Nebraska, Montana, Massachusetts, Kentucky, Kansas, Iowa, Hawaii, and Idaho. The treatment states are those in the required retention category for third grade under state law and have at least

2 post-period observations within my data window (2010-2019)[3]. Whereas in model 3, I consider control states that allow for retention but are not required by states: Texas, Colorado, Minnesota, West Virginia, New Mexico, and Maryland (Cummings & Turner, 2020). It is worth noting that, compared to my treatment states, the control states in model 3 also include diagnostic/screening assessments and interventions to improve reading. As the policy is state-level, I cluster the standard errors at the state level to allow arbitrary within-state error correlation.

Because negative weighting problems arise when using the canonical TWFE (Two-Way Fixed Effects) models with heterogeneous treatment timing, I use the approach of Callaway and Sant'Anna (2021) to address this underestimation. Identification comes from variation over time and across cohorts. I use doubly robust augmented inverse-probability weighting (AIPW) to estimate the ATETs.

I observe data $\{y_{it}, \mathbf{x}_{it}, d_{it}\}$ for unit i at time t with unit i, $i = 1,\ldots, N$, and time t, with t = 1,…, T. Where y_{it} is the observed outcome (RLA Score), x_{it} are covariates for the outcome model, and d_{it} is an indicator that equals one if an observation is treated or zero otherwise.

Let $Y_{it}(0)$ denote the untreated potential outcome of the unit i at time t, and $Y_{it}(g)$ the potential outcome if the unit i becomes treated in the period g. Each unit belongs to the treatment-timing group G_i(the first period of treatment).

$$Y_{it} = (1\{G_i > t\})Y_{it}(0) + (1\{G_i \leq t\})Y_{it}(G_i),$$

The equation suggests that untreated potential outcomes are observed for not-yet-treated units, and post-treatment units have treated potential outcomes. This methodology uses no-anticipation and parallel-trends assumptions to identify group-time average treatment effects $ATET(g, t)$, and combine them into an overall ATET using weights. In my educational reform policy contexts, the adoption of treatment is staggered, and once a unit is treated, it remains treated in all consecutive periods ($D_{it} = 1 \Rightarrow D_{i,t+1} = 1$).

Identification relies on a parallel trends assumption. I can formulate the assumption based on the not-yet-treated group.

$$E[Y_t(0) - Y_{t-1}(0) \mid G = g] = E[Y_t(0) - Y_{t-1}(0) \mid D_s = 0, G \neq g],$$

Which uses units that have not been treated by time $s \geq t$ as valid controls, this method uses more information, although it makes it harder to see pre-treatment trends across groups (Marcus & Sant'Anna, 2021).

[3] Here, Model 1 considers the treatment states, starting from the first grade 4 cohort facing the retention policy in their third grade. While Model 2 establishes an anticipation window for Mississippi, considering its unique preparation to improve students' early reading skills. Finally, Model 3 considers the same condition as Model 1, but now the control states allow retention, but are not required by the states.

Here, my parallel trend assumption for the models is subject to pre-treatment covariates. This results in the conditional parallel trends assumption, which allows for covariate-specific variations in outcome dynamics across groups.

$$E[Y_t(0) - Y_{t-1}(0) \mid X, G = g] = E[Y_t(0) - Y_{t-1}(0) \mid X, D_s = 0, G \neq g].$$

These conditional analogs of the standard parallel trends assumptions allow heterogeneous pre-treatment trajectories influenced by observed covariates X, thereby increasing the feasibility of identification in contexts where the distribution of covariates varies across treatment groups.

The preceding assumptions inherently broaden the identifying conditions from the two-period, two-group DiD framework to contexts with multiple periods. Similarly, the treatment parameter of interest can be generalized by establishing group-time average treatment effects as

$$ATET(g,t) = E[Y_t(g) - Y_t(0) \mid X, G = g],$$

This shows the average treatment effect for the group that was first treated in period g at time t. For instance, in a three-period design, $ATET$ (g=2, t=3) shows the effect for units that were treated in period 2 and measured in period 3.

For *not-yet-treated* comparison groups, it becomes

$$ATET(g,t) = E[Y_t - Y_{g-1} \mid X, G = g] - E[Y_t - Y_{g-1} \mid X, D_s = 0, G \neq g].$$

These $ATET(g,t)$ estimates serve as the fundamental building blocks for understanding treatment effects in Difference-in-Differences designs with staggered adoption and multiple time periods. In Difference-in-Differences designs with more than one group and period, group-time average treatment effects (ATET(g,t)) are the natural parameters because they capture how treatment effects vary across both dimensions. To facilitate straightforward interpretation, I aggregated these estimates in my analysis and computed group-specific and overall average treatment effects for the treated.

Following Callaway and Sant'Anna (2021)[4] one useful summary is the group-specific average effect:

[4] The analysis was conducted in STATA 19.5 BE using *xthdidregress aipw* command.

$$\theta_S(g) = \frac{1}{T-g+1}\sum_{t=2}^{T} 1\{g \le t\}ATET(g,t),$$

This calculates the mean treatment effect for units initially treated in period g. This parameter shows the average of all *post-treatment* effects for the cohort that was first treated in the period g.

Then, an easy-to-understand summary is

$$\theta_{OS} = \sum_{g=2}^{T} \theta_S(g)P(G = g \mid G \le T),$$

This is like the ATET in the standard two-period case and shows the overall average treatment effect across all treated groups. Where, $Pr(G = g \mid G \le T)$ represents the share of states first treated in the year g among all states that received the policy within the observed sample period. In practice, this probability is obtained by dividing the number of states whose first treatment year equals g by the total number of treated states observed up to year T.

I also examine treatment effect dynamics, that is, the temporal evolution of effects. This can be recorded by

$$\theta_D(e) = \sum_{g=2}^{T} 1\{g + e \le T\}ATET(g, g+e)P(G = g \mid G + e \le T),$$

Which shows the average effect for groups that have been treated for exactly e periods. This aggregation is similar to event-study analyses, which provide a dynamic view of how treatment effects change over time. The probability term $Pr(G = g \mid G + e \le T)$ is an empirical weighting factor, representing the share of treated states whose first treatment year equals g among all cohorts that can be observed e years after treatment. In practice, this is computed by counting how many states were first treated in each g and are still in the sample at the time $g + e$, divided by the total number of observable treated cohorts for that event time. These probabilities, therefore, ensure that the aggregated dynamic effect $\theta_D(e)$ reflects the correct composition of cohorts for each exposure period, giving a meaningful average effect aligned with the duration of exposure rather than calendar year.

3.1 Data

This paper used district-level grade 4 RLA scores and covariate data from SEDA (Stanford Education Data Archive). The SEDA dataset links student performance on state-specific assessments by standardizing results relative to the National Assessment of Educational Progress (NAEP). In the dataset, there are 490,000 district grade observations for RLA (Bleiberg et al., 2025). In each grade-subject, identical national means and standard deviations are used to standardize estimated scores across all years, ensuring that variations in standardized scores over time reflect absolute changes in national standard deviation units rather than relative changes (Matheny et al., 2023).

Districts are significant entities because of their substantial variations in sociodemographic composition and their authority over numerous educational policies (budget distribution, curriculum, disciplinary regulations). The National Assessment of Educational Progress (NAEP) is a consistent, nationally representative examination conducted biennially for fourth- and 8th-graders. SEDA correlates state exam scores with NAEP scores, thereby establishing a uniform, consistent scale for all state and district test outcomes. In the absence of this linkage, trend studies could be deceptive; an apparent gain may really result from the implementation of a less challenging assessment. Through NAEP-based scaling, researchers may distinguish genuine achievement trends from artifacts resulting from changes in state testing practices.

Table 2: Descriptive Statistics of the Study Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
Mean RLA Score for all students (RLA_ALL)	39224	.117	.362	-2.749	1.599
Unemployment rate (UR)	39467	.067	.027	0	.285
Log of median income (LMI)	39467	10.858	.305	9.58	12.391
Urbanicity (UR)	.	.	.	.	.
City	39561	.05	.218	0	1
Suburb	39561	.242	.428	0	1
Town	39561	.219	.414	0	1
Rural	39561	.489	.5	0	1
Snap Receipt Rate (SR)	39467	.106	.066	0	.488
Percent Free and Reduced Lunch (PFRL)	39557	.457	.214	0	1

Table 2 displays the summary statistics of my primary variables of interest. Here, covariates are based on education policy research examining the impact on achievement scores (Shores & Steinberg, 2017; Ruffini, 2020; Bleiberg et al., 2025; Bellows, 2019). Here, the dependent variable is the mean RLA score for all students, which averages 0.117, suggesting the mean score is above the national average, while values range from -2.749 to 1.599; thus, variability in academic performance exists across different districts and years.

All other variables listed are control variables for my DiD models; the descriptive characteristics of these variables suggest that the sample is skewed towards less urbanized regions. The unemployment rate (UR) averages 6.7% with slight variation (std. dev. = 0.027), while the log of median income (LMI) averages 10.86 with greater variation (std. dev. = 0.305). Among the socioeconomic covariates, the SNAP receipt rate (SR) averages 10.6%, and the percentage of students eligible for free and reduced-price lunch (PFRL) averages 45.7%, indicating moderate to high levels of poverty exposure across the sampled districts.

4.0 Results

4.1 Considering no retention policy as the control group (Model 1)

Table 3: Tabulation of treatment state observation (no anticipation)

	Spring of Tested Year					
State Abbreviation	2015	2016	2017	2018	2019	Total
AZ	144	130	132	83	146	635
MS	0	129	139	138	139	545
NC	115	115	115	115	118	578
OH	565	600	604	605	605	2979
Total	824	974	990	941	1008	4737

As shown in the table, except for MS, treatment began in 2015, while in Mississippi, it began in 2016. This start date aligns with the first cohort to receive this intervention, which is in grade 4. The values for each state over time indicate that I have a consistent number of observations throughout the period. This table is created on the assumption that there is no anticipation of any state.

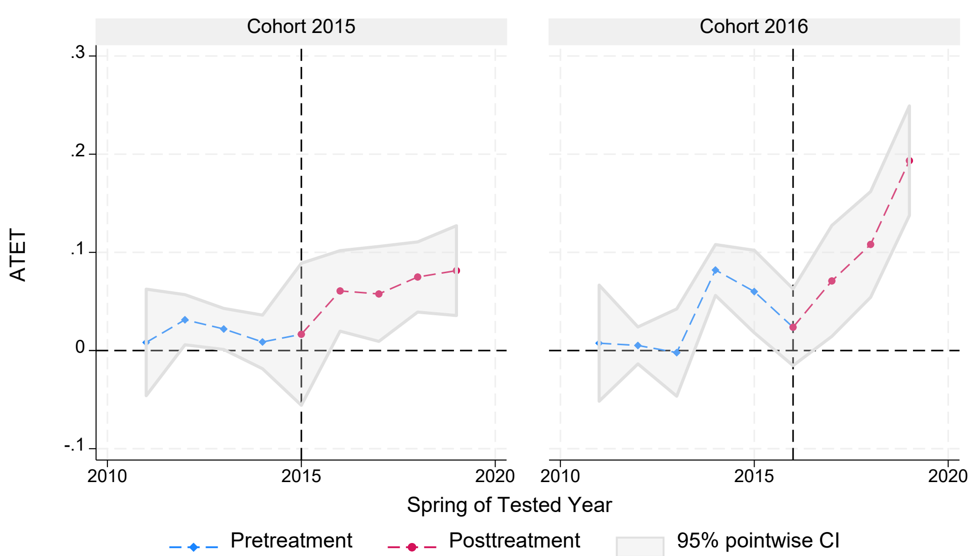

Figure 1. Dynamic Average Treatment Effects on the Treated (ATETs) by Cohort, 2010–2019
Estimated using Augmented Inverse Probability Weighting (AIPW); outcome: district mean 4th-grade RLA score (SEDA)

Using the untreated comparison group and district-level covariates (unemployment, log median income, urbanicity, SNAP receipt, PFRPL), I calculate cohort-specific dynamic Average Treatment Effects on the Treated (ATETs) for the SEDA RLA score from 2011 to 2019 for two treated cohorts (first treated in 2015, then in 2016). As shown in Figure 1 and Table A1, for the 2015 cohort, the majority of pre-treatment leads (2011–2014) are minimal and statistically indistinguishable from zero, indicating parallel pre-trends for this cohort; effects become positive post-treatment and increase over time, becoming statistically significant by 2016–2019. For the 2016 cohort, two years prior to retention treatment, ATETs are already positive and statistically significant, suggesting some anticipation or early gains or a mild violation of parallel pre-trends for this cohort; post-treatment effects are consistently positive and statistically significant, rising through 2017–2019 and reaching about +0.193 SD in 2019. Both cohorts exhibit significant enhancements in RLA achievement post-policy adoption, with the 2016 cohort demonstrating more substantial and rapid progress. Methodologically, since the base time for pretreatment ATETs is the previous period, insignificant pre-coefficients suggest no differential changes immediately prior to treatment; the significant 2014–2015 leads for the 2016 cohort necessitate a robustness check (e.g., incorporating an anticipation window or redefining the cohort start).

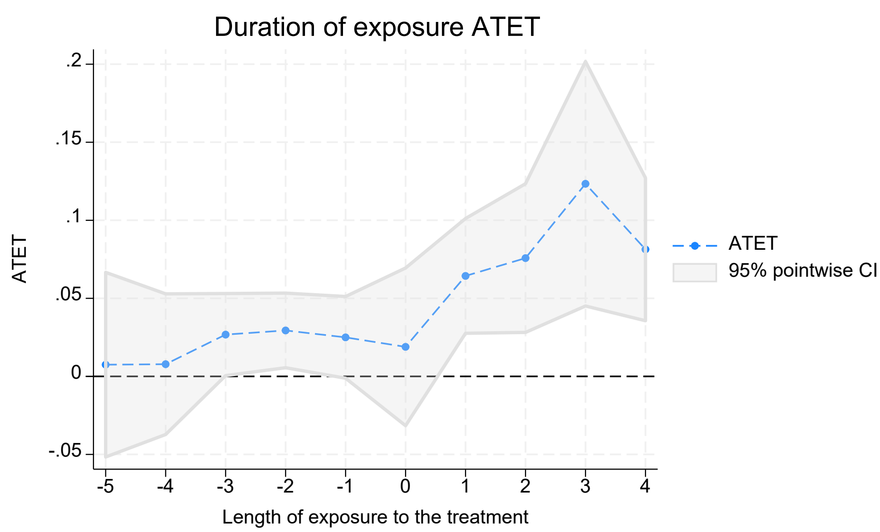

Figure 2. Event Study by Duration of Exposure to the Policy

Table A2 and *Figure 2* show dynamic ATETs, summed across treated groups, by the length of exposure to the policy or by the number of years since the policy was first implemented. The pre-treatment coefficients (exposures -5 to -1) are small and mostly not statistically significant, except for exposures -3 and -2, which suggest a slight increase in RLA achievement three years before the policy was put into place. While this may suggest limited anticipation or pre-trend disparities, the overall pre-policy trend remains stable around zero, providing conditional support for the parallel-trends hypothesis. After treatment (exposures ≥ 0), ATETs consistently yield positive, statistically significant results, with effect sizes escalating over time. The estimated effect grows from 0.018 SD in the year after adoption (not significant) to 0.081 SD four years later ($p < 0.01$). This monotonic post-treatment trajectory suggests that the reading-retention reforms led to ongoing improvements in district-level RLA performance over time.

Table 4 Overall and Cohort-Specific Average Treatment Effects on the Treated (ATETs)

ATET	**AIPW Estimates**
Overall	.071** (.019)
Cohort 2015	.057** (.018)
Cohort 2016	.100** (.024)
Number of Observations	39,132

Note: ATET aggregate and robust standard error in parentheses
** indicates significance at 1% level

In *Table 4*, I see three significant ATET estimates. The first estimate is the overall ATET, which is the average across all treated cohorts and years. The estimated impact of the third-grade reading retention policy on mean RLA achievement is 0.071 standard deviations, indicating a statistically significant and educationally relevant improvement in literacy performance in treated districts relative to untreated counterparts. The other two break down the results by treatment group. The 2015 adopters had an average post-treatment gain of 0.057 SD, while the 2016 adopters had a bigger effect of 0.100 SD.

4.2 Considering anticipation for Mississippi (Model 2)

After facing issues with the parallel pre-trend assumption of Mississippi in model 1, I am considering the anticipation window.

Table 5: Tabulation of treatment state observation (Anticipation for MS)

	Spring of Tested Year							
State Abbreviation	2013	2014	2015	2016	2017	2018	2019	Total
AZ	0	0	144	130	132	83	146	635
MS	138	124	135	129	139	138	139	942
NC	0	0	115	115	115	115	118	578
OH	0	0	565	600	604	605	605	2979
Total	138	124	959	974	990	941	1008	5134

Table 5 suggests an allocation of treated observations similar to *Table 3*; the observation over the year is the same across all states except Mississippi. In this model, I assume that Mississippi may experience preliminary impacts of the third-grade retention policy before its official start, suggesting that schools and educators may have altered their teaching methods after the policy was announced.

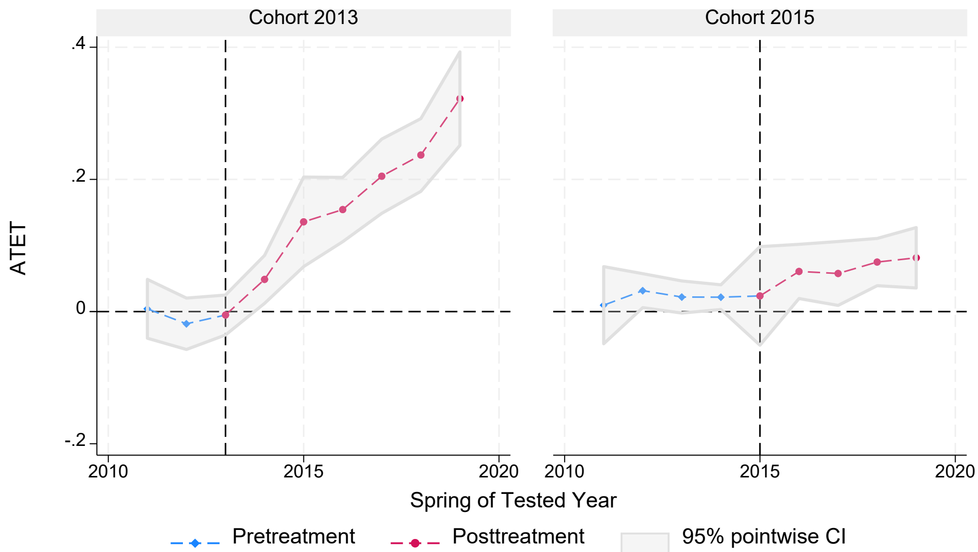

Figure 3. Dynamic Average Treatment Effects on the Treated (ATETs) by Cohort, 2010–2019

Table A3 and Figure 3 illustrate the progression of the effects of third-grade retention policies over time for the 2013 and 2015 adoption cohorts. For the 2013 cohort (Mississippi), all pre-treatment estimates (2011–2012) are minimal and statistically insignificant, indicating that treated and control districts displayed parallel trends prior to policy implementation. The ATET becomes positive and statistically significant (0.049, $p<0.01$) in 2014, the first year after the treatment. The effect increases steadily over subsequent years, reaching 0.136 in 2015, 0.154 in 2016, 0.205 in 2017, 0.237 in 2018, and 0.322 in 2019. These results suggest a cumulative enhancement in RLA performance within the treated districts, indicating that the policy's effect intensifies over time as exposure increases. The same thing seems to be happening with the 2015 cohort: almost all the pre-treatment coefficients (2011–2014) are small and not statistically significant, which supports the idea that there was no pre-policy divergence. Beginning in 2016, the first full year of treatment, ATETs became positive and statistically significant. They start at 0.024 in 2015 and rise each year to 0.06 in 2016, 0.058 in 2017, 0.075 in 2018, and 0.081 in 2019. This progression indicates that the policy's efficacy was not instantaneous but increased as the program developed, likely due to improvements in implementation quality, instructional modifications, and greater familiarity among educators and administrators.

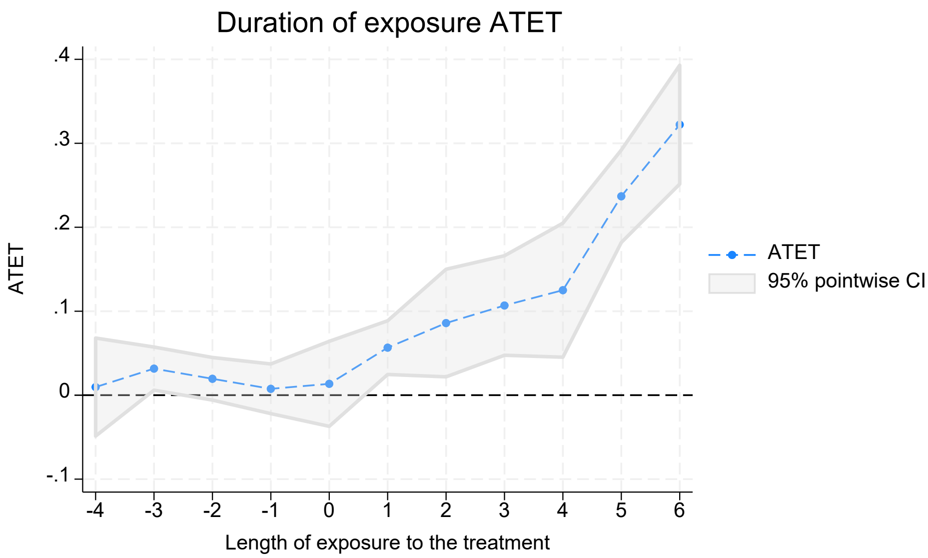

Figure 4. Average Treatment Effects by Duration of Exposure to the Third-Grade Retention Policy

In this case (*Figure 4 and Table A4*), the variable Exposure is the number of years since a state first implemented the policy. Negative values indicate years before treatment, and 0 indicates the first year of treatment. The findings show that the pre-treatment estimates (exposure = –4 to –1) are small and not statistically significant, which means that there were no noticeable differences between the treated and control districts before the policy was put into place.

Post-treatment estimates (exposure ≥ 1), on the other hand, are always positive and statistically significant. This suggests that the policy's effects get stronger over time. The estimated ATET increases from 0.014 in the first year of implementation to 0.086 after 2 years, 0.107 after 3 years, 0.125 after 4 years, and 0.322 after 6 years of exposure. This gradual rise in treatment effects indicates that the retention policy is having a cumulative, more substantial effect on district-level RLA achievement as districts are exposed to the intervention for longer periods. The pattern indicates that the advantages of implementation are likely to emerge incrementally as schools modify instructional methods, offer remediation, and assimilate policy-driven accountability systems.

Table 6: Overall and Cohort-Specific Average Treatment Effects on the Treated (ATETs)

ATET	AIPW Estimates
Overall	.103** (.035)
Cohort 2013	.159** (.022)
Cohort 2015	.059** (.018)
Number of Observations	39,132

Note: ATET aggregate and robust standard error in parentheses
** indicates significance at 1% and 5% level

As shown in *Table 6*, the estimated ATET for Mississippi (2013 Cohort) is 0.159 ($p < 0.01$), which is higher than the two other estimates. This could be due to changes in how teachers teach, more teacher training, or a comparatively greater focus on literacy because of the new policy.

The 2015 cohort marks the actual implementation period for other adopting states (excluding Mississippi), during which the policy was fully operational. The estimated ATET for this group is 0.059 ($p < 0.01$), which means that RLA performance improved significantly after adoption. At the same time, the overall ATET for all treated districts is 0.103 ($p < 0.01$). This means that districts exposed to third-grade retention policies did, on average, better in RLA than districts that had not yet been treated. In conclusion, these results indicate that the policy had measurable academic benefits, consistent with Model 1's findings.

4.3 Considering control states where retention is optional (Model 3)

Here, I am keeping treatment states and treatment timing the same as model 1. My control states that have state intervention for improving reading and assessment of reading skills, without the state-based retention mandate; that is, they are leaving the decision to the local level. So, the goal is to see the impact of state-mandated retention compared to states where it is left to districts.

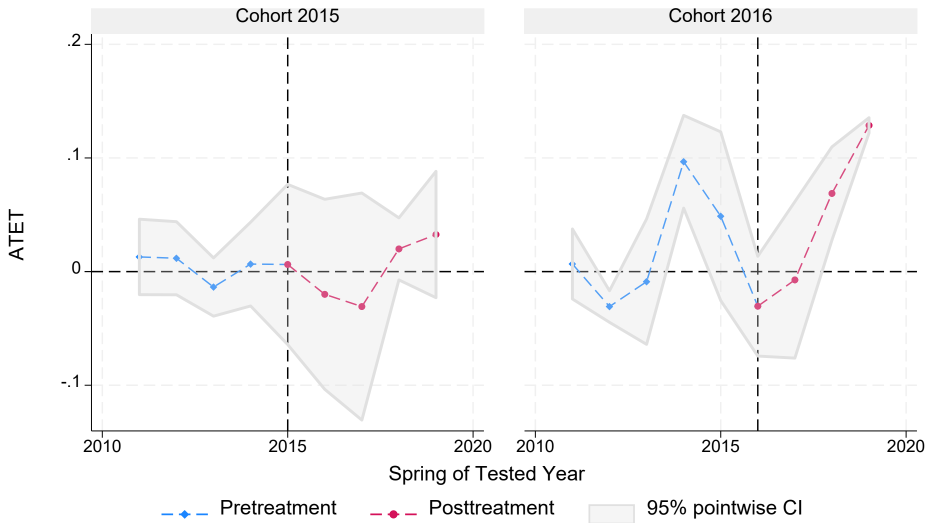

Figure 5. Dynamic Average Treatment Effects on the Treated (ATETs) by Cohort, 2010–2019

As shown in Table A5 and Figure 5, the pre-treatment estimates for the 2015 cohort (2011–2014) are minimal and statistically insignificant (p-values range from 0.42 to 0.73), indicating that treated and control districts exhibited parallel trends before the policy was implemented. In the post-treatment period, effects are statistically insignificant, suggesting no impact of the policy. In the pre-treatment period for the 2016 cohort, several significant changes are observed: a positive change in 2014 ($p<0.01$) and a negative change in 2012 (−0.031, $p<0.01$). This suggests that there may have been violations of the parallel pre-trend or no anticipation assumption before formal adoption. After implementation, the ATET became positive, with statistically significant improvements seen in 2018 (0.069) and 2019 (0.106). These results indicate that, despite initial instability in districts prior to adoption, the policy's beneficial effects emerged a few years post-implementation and intensified over time.

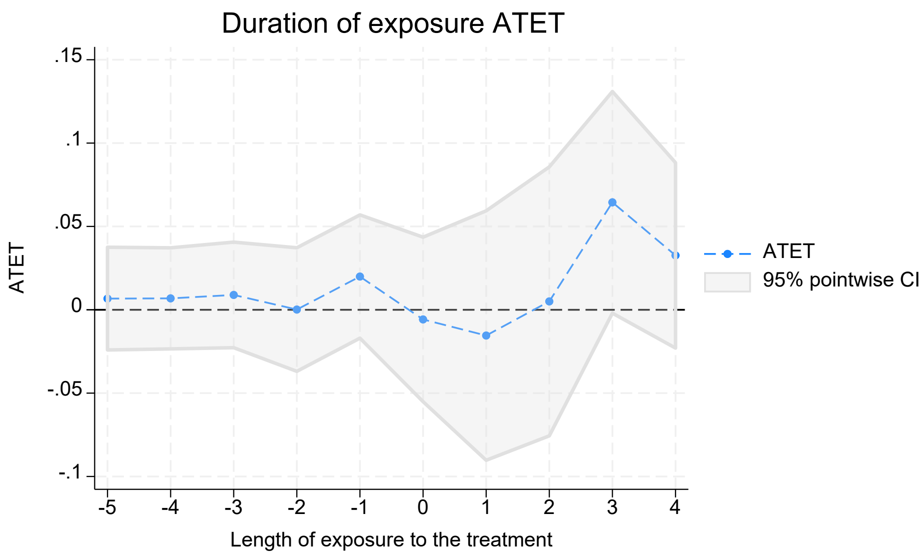

Figure 6 Average Treatment Effects on the Treated (ATETs) by Duration of Exposure

The exposure variable counts the number of years since the first treatment year. Negative values indicate years before treatment, and 0 indicates the year of adoption. As per *Table A6* and *Figure 6*, during the pre-treatment period (exposure −5 to −1), the coefficients are minimal and statistically insignificant, suggesting that treated and control districts exhibited parallel trends prior to implementation. At exposure 0 (the first year of treatment), the effect is still not significant (−0.0057, p = 0.819), indicating no immediate improvement in RLA performance at the district level. In the years following treatment, the ATETs remain nearly zero during the initial two periods (exposure 1–2). Although the estimated coefficients exhibit slight positive values at extended durations, 0.064 (p = 0.057) after three years and 0.033 (p = 0.250) after four years, the effects do not consistently achieve statistical significance.

Table 7: Overall and Cohort-Specific Average Treatment Effects on the Treated (ATETs)

ATET	**AIPW Estimates**
Overall	.014
	(.019)
Cohort 2015	.001
	(.023)
Cohort 2016	.041**
	(.011)
Number of Observations	14,224

Note: ATET aggregate and robust standard error in parentheses
** indicates significance at 1% and 5% level

As per *Table 7*, the overall ATET estimate is small and not statistically significant (0.014, $p = 0.470$). This means that when I look at all the states that were treated, I cannot see any average effect of the third-grade retention policy on 4th-grade RLA achievement at the district level. In the cohort-specific analysis, the 2015 cohort shows a negligible, statistically insignificant effect (0.001, $p = 0.950$), indicating that districts in AZ, NC, and OH did not achieve measurable improvements in RLA performance compared to untreated districts. The 2016 cohort (MS), on the other hand, shows a positive and statistically significant average effect (0.041, $p < 0.01$), indicating that districts in MS experienced a small but significant improvement of about 0.04 standard deviations in mean RLA scores. These findings indicate that, on average, solely the third-grade retention aspect of the policy yielded minimal or negligible effects across all treated districts of AZ, NC, and OH, while an exception exists for MS.

5.0 Discussion

Table 8: Comparison of ATET estimates among three models

ATET	1 Considering no retention policy control	2 Considering MS Anticipation	3 Isolating State-Mandated Retention
Overall	.071** (.019)	.103** (.035)	.014 (.019)
Cohort 2015 (AZ, OH, and NC)	.057** (.018)	.059** (.018)	.001 (.023)
Cohort MS *	.100** (.024)	.159** (.022)	.0.041** (.011)
Number of Observations	39,132	39,132	14,224

Note: For MS in Models 1 and 3, the 2016 cohort is used, while for Model 2, MS is the 2013 cohort. The robust standard error is in parentheses. Moreover, ** indicates significance at the 1% level.

Table 8 presents the Average Treatment Effects on the Treated (ATET) for three model specifications that examine how third-grade retention policy affects fourth-grade reading skills. Model 1 considers the treatment states, starting from the first grade 4 cohort facing the retention policy in their third grade, with no anticipation assumption. In model 1, the overall estimated effect is 0.071 ($p<0.05$). This means that, on average, exposure to a state retention policy raised mean RLA scores by about 0.07 standard deviations. When anticipation effects for Mississippi are included in the model (Model 2), the overall ATET increases to 0.103 ($p<0.05$). This means that including the years before the policy officially went into effect (2013–2015) for Mississippi shows extra gains. Model 3, on the other hand, examines the effect of state-mandated retention by comparing states with a required retention policy to those with literacy laws that allow districts to decide whether to retain students. This model shows a small and statistically insignificant overall effect (0.014). This pattern suggests that the positive effects observed earlier are not solely due to the mandate itself, but rather to the broader range of literacy supports that accompany policy adoption. Overall, in my study, I observe ATET in the range of 0.07 to 0.10 in standard deviation gain, which is supported by recent studies on the effects of early literacy policy by Westall and Cummings (2023) and Spencer (2024).

The 2015 group (Arizona, Ohio, and North Carolina) shows consistent and significant gains of about 0.058–0.059 standard deviations in Models 1 and 2. The effect vanishes in Model 3, indicating that the retention state mandate alone does not account for the observed improvement when comparing states with analogous literacy initiatives (Goos et al., 2021). Mississippi's performance is the most impressive: its estimated effect is 0.100 in Model 1, 0.159 in Model 2, and remains positive and significant (0.041) in Model 3. This shows that Mississippi's gains are both significant and robust enough to withstand different definitions of control. This is reiterated by Spencer (2024), who focused on Mississippi's early literacy case and found a greater policy impact compared to other states.

In 2013, Mississippi was placed 49th among the 50 states on the NAEP Grade 4 reading assessment. In response to inadequate performance, state legislators enacted the Literacy-Based Promotion Act (LBPA) in April 2013 to ensure that all Mississippi students achieve reading proficiency by the end of third grade. The Act allocates up to $15 million per year for literacy efforts, structured around three primary pillars: enhancing instruction, identifying and supporting students with reading deficiencies, and retaining third graders who do not meet reading criteria (Huebeck, 2023; Spencer, 2024). Further, the LBPA mandated that the Mississippi Department of Education (MDE) initiate a statewide teacher training program in January 2014, focusing on scientifically based reading instruction (Folsom et al., 2017). The program additionally implemented literacy coaches to support underperforming schools. MDE initially employed 29 coaches in 2013–2014 and increased this number to 74 highly competent coaches across 126 schools by 2015–2016 (Howard-Brown et al., 2016). So, I would say the consistent positive impact of the retention policy is firmly attributable to its diversified interventions; that is, in Mississippi, teacher training and coaching are a significant distinguishing feature of the policy, not retention itself.

6.0 Conclusion

This study adds to the growing body of research on early literacy policy by showing how state-mandated third-grade retention affects students differently from other reading interventions. Employing a staggered adoption Difference-in-Differences framework and a decade of district-level achievement data, the analysis reveals modest yet significant improvements in fourth-grade RLA performance after the introduction of retention mandates. However, when examining the mandate in isolation compared to states with similar literacy initiatives that allow local discretion over retention, these beneficial effects diminish. The results indicate that the mandate alone will not improve literacy outcomes; instead, success depends on how well the policies are implemented and on the resources available for teaching. Mississippi's experience is a clear example of this difference. The Literacy-Based Promotion Act (LBPA) in the state not only required retention for students who were not meeting the minimum reading threshold, but also allocated resources to teacher training, literacy coaching, and early interventions for students in grades K–3. These elements led to enduring enhancements in student performance, aligning with earlier research by Spencer (2024) and Westall and Cummings (2023), which illustrates the importance of comprehensive, multi-tiered literacy policies in facilitating educational achievement.

The results generally indicate that the process of formulating a policy holds greater significance than its actual content. The state-mandated retention policy only works as it comes with the literacy interventions needed to keep students from being retained in the first place. Therefore, to enhance students' early-grade reading skills, policymakers should prioritize providing teachers with the necessary tools, early diagnostic systems, and long-term reading interventions that address the root causes of students' poor reading performance rather than merely treating the symptoms. The findings also suggest that state policymakers should be cautious about imposing additional retention mandates when district-level discretion and literacy interventions already function effectively.

There can be discrepancies in the strength of retention policy based upon how many students are retained in each state; this paper did not account for this, that is, it did not use student-level data to show whether retention works. On the contrary, it aims to assess the efficacy of state-mandated retention as a policy using district-level 4th-grade RLA scores. Furthermore, the study concentrates on short-term outcomes, specifically fourth-grade RLA scores in the years immediately following policy implementation. It cannot account for long-term effects, such as retention's impact on middle school performance, dropout rates, or high school graduation. Previous studies (e.g., Schwerdt et al., 2017; Jacob & Lefgren, 2009) indicate that short-term gains may diminish; thus, these findings should not be construed as indicators of enduring enhancements in achievement. Overall, the paper offers reliable short-term evidence regarding state-mandated retention policies; however, it is constrained by its aggregate nature, brief observation period, possible confounding policy overlap, and the distinctive efficacy of Mississippi's intervention. Subsequent research employing student-level longitudinal data and multi-state policy audits may elucidate the mechanisms.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Angrist, J. D., & Lavy, V. (1998). *Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools* (No. w6781). National Bureau of Economic Research. https://doi.org/10.3386/w6781

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bassi, M., Meghir, C., & Reynoso, A. (2020). Education Quality and Teaching Practices. *The Economic Journal*, *130*(631), 1937–1965. https://doi.org/10.1093/ej/ueaa022

Berne, Jordan S., Jacob, Brian A., Weiland, Christina, & Strunk, Katharine O. (2025). The Impacts of Grade Retention Policy With Minimal Retention. *EdWorkingPapers*. https://doi.org/10.26300/TQRB-YH12

Bellows, L. (2019). Immigration Enforcement and Student Achievement in the Wake of Secure Communities. *AERA Open*, *5*(4), 2332858419884891. https://doi.org/10.1177/2332858419884891

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (2025). Taking Teacher Evaluation to Scale: The Effect of State Reforms on Achievement and Attainment. *Journal of Political Economy Microeconomics*, *3*(3), 568–610. https://doi.org/10.1086/732837

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press.

Churchill, A. (2023). *More evidence—This time from Ohio—That third grade retention works*. The Thomas B. Fordham Institute. https://fordhaminstitute.org/ohio/commentary/more-evidence-time-ohio-third-grade-retention-works

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Cilliers, J., Fleisch, B., Prinsloo, C., & Taylor, S. (2019). How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources*. https://doi.org/10.3368/jhr.55.3.0618-9538R1

Cummings, A., & Turner, M. (2020). *COVID-19 and Third-Grade Reading Policies:* EPIC.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*(6), 934–945. https://doi.org/10.1037/0012-1649.33.6.934

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

DiSalvo, R. W., & Yu, J. H. (2023). Housing Affordability and School Quality in the United States. *Journal of Housing Economics*, *60*, 101933. https://doi.org/10.1016/j.jhe.2023.101933

Duke, N., Moje, E., & Palincsar, A. (n.d.). *Three IRA Literacy Research Panel Members Comment on Michigan House Bill 5111*. Retrieved November 3, 2025, from https://www.literacyworldwide.org/blog/literacy-now/2014/02/26/three-ira-literacy-research-panel-members-comment-on-michigan-house-bill-5111

ExelinEd. (2021). *Comprehensive early literacy policy: State-by-state analysis of fundamental principles*.

Feng, L., & Sass, T. R. (2013). What makes special-education teachers special? Teacher training and achievement of students with disabilities. *Economics of Education Review*, *36*, 122–134. https://doi.org/10.1016/j.econedurev.2013.06.006

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Fiester, L. (2013). *Early warning confirmed: A research update on third-grade reading*. The Annie E. Casey Foundation.

Folsom, J. S., Smith, K. G., Burk, K., & Oakley, N. (2017). *Educator Outcomes Associated with Implementation of Mississippi's K-3 Early Literacy Professional Development Initiative. REL 2017-270*. Regional Educational Laboratory Southeast. https://eric.ed.gov/?id=ED573545

García-Pérez, J. I., Hidalgo-Hidalgo, M., & Robles-Zurita, J. A. (2014). Does grade retention affect students' achievement? Some evidence from Spain. *Applied Economics*, *46*(12), 1373–1392. https://doi.org/10.1080/00036846.2013.872761

Goos, M., Pipa, J., & Peixoto, F. (2021). Effectiveness of grade retention: A systematic review and meta-analysis. *Educational Research Review*, *34*, 100401. https://doi.org/10.1016/j.edurev.2021.100401

Greene, J. P., & Winters, M. A. (2007). Revisiting Grade Retention: An Evaluation of Florida's Test-Based Promotion Policy. *Education Finance and Policy*, *2*(4), 319–340. https://doi.org/10.1162/edfp.2007.2.4.319

Greene, J. P., & Winters, M. A. (2009). The effects of exemptions to Florida's test-based promotion policy: Who is retained?: Who benefits academically? *Economics of Education Review*, *28*(1), 135–142. https://doi.org/10.1016/j.econedurev.2008.02.002

Hanushek, E. A., & Woessmann, L. (2012). Schooling, educational achievement, and the Latin American growth puzzle. *Journal of Development Economics*, *99*(2), 497–512. https://doi.org/10.1016/j.jdeveco.2012.06.004

Hernandez, D. J. (2011). Double Jeopardy: How Third-Grade Reading Skills and Poverty Influence High School Graduation. In *Annie E. Casey Foundation*. Annie E. https://eric.ed.gov/?id=ED518818

Heubeck, E. (2023). Mississippi Students Surged in Reading Over the Last Decade. Here's How Schools Got Them There; Educators have chalked it up to a multipronged effort, spurred by the state's literacy act. *Education Week*, *42*(37), NA-NA.

Ho, A. D. (2020). What Is the Stanford Education Data Archive Teaching Us About National Educational Achievement? *AERA Open*, *6*(3), 2332858420939848. https://doi.org/10.1177/2332858420939848

Jacob, B. A., & Lefgren, L. (2009). The Effect of Grade Retention on High School Completion. *American Economic Journal: Applied Economics*, *1*(3), 33–58. https://doi.org/10.1257/app.1.3.33

James, J. (2024, May 22). *How many students are retained by the 'third-grade gate'? No one knows - Mississippi Today*. https://mississippitoday.org/2024/05/22/how-many-students-are-retained-by-the-third-grade-gate-no-one-knows/

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*, *88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Loewus, L. (2015, October 16). 3rd Grade Reading Retention: A Closer Look at North Carolina's Literacy Law. *Education Week*. https://www.edweek.org/teaching-learning/3rd-grade-reading-retention-a-closer-look-at-north-carolinas-literacy-law/2015/10

Manacorda, M. (2012). The Cost of Grade Retention. *The Review of Economics and Statistics*, *94*(2), 596–606. https://doi.org/10.1162/REST_a_00165

Manson, S., Schroeder, J., Van Riper, D., & Ruggles, S. (2019). *SEDA File: Version 14.0* (Version 14.0) [Dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D050.V14.0

Matheny, K. T., Thompson, M. E., Townley-Flores, C., & reardon, sean f. (2023). Uneven Progress: Recent Trends in Academic Performance Among US School Districts. *American Educational Research Journal*, *60*(3), 447–485. https://doi.org/10.3102/00028312221134769

Özek, U. (2015). Hold Back To Move Forward? Early Grade Retention And Student Misbehavior: *Education Finance and Policy*, *10*(3), 350–377. https://doi.org/10.1162/EDFP_a_00166

Planty, M., Hussar, W., Snyder, T., Provasnik, S., Kena, G., Dinkes, R., KewalRamani, A., & Kemp, J. (2008). *The Condition of Education, 2008. NCES 2008-031*. National Center for Education Statistics. https://eric.ed.gov/?id=ED501487

Perrault, P., & Winters, M. A. (2020, July 28). Test-Based Promotion and Student Performance in Florida and Arizona. *Manhattan Institute*. https://manhattan.institute/article/test-based-promotion-and-student-performance-in-florida-and-arizona

Read On Arizona. (2018). 70 Percent of 4th-Graders Don't Read Proficiently—What's the Problem? *Read On Arizona*. https://readonarizona.org/news/70-percent-of-4th-graders-dont-read-proficiently-whats-the-problem/

Read On Arizona. (2025). *Move On When Reading*. https://readonarizona.org/resources/move-on-when-reading/

Ruffini, K. (2020). *Universal Access to Free School Meals and Student Achievement: Evidence from the Community Eligibility Provision*.

Schwerdt, G., West, M. R., & Winters, M. A. (2017a). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*, 154–169. https://doi.org/10.1016/j.jpubeco.2017.06.004

Schwerdt, G., West, M. R., & Winters, M. A. (2017b). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*, 154–169. https://doi.org/10.1016/j.jpubeco.2017.06.004

Schwerdt, G., West, M. R., & Winters, M. A. (2017c). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*(C), 154–169.

Shores, K., & Steinberg, M. (2017). The Impact of the Great Recession on Student Achievement: Evidence from Population Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3026151

Sparks, R. L., Patton, J., & Murdoch, A. (2014). Early reading success and its relationship to reading achievement and reading volume: Replication of '10 years later.' *Reading and Writing*, *27*(1), 189–211. https://doi.org/10.1007/s11145-013-9439-2

Spencer, N. (2024a). Comprehensive early literacy policy and the "Mississippi Miracle." *Economics of Education Review*, *103*, 102598. https://doi.org/10.1016/j.econedurev.2024.102598

Spencer, N. (2024b). Comprehensive early literacy policy and the "Mississippi Miracle." *Economics of Education Review*, *103*, 102598. https://doi.org/10.1016/j.econedurev.2024.102598

Wassmer, R. W., & Turgeon, M. D. (2024). Particulates Matter: The Influence of Cumulative Local Air Pollution Exposure on Sixth-Grade Academic Achievement in California. *Environmental Health Insights*, *18*, 11786302241293292. https://doi.org/10.1177/11786302241293292

Westall, J., & Cummings, A. (2023). *The Effects of Early Literacy Policies on Student Achievement*. Education Policy Innovation Collaborative. https://eric.ed.gov/?id=ED672706

Winters, M. A., & Greene, J. P. (2012). The Medium-Run Effects of Florida's Test-Based Promotion Policy. *Education Finance and Policy*, *7*(3), 305–330. https://doi.org/10.1162/EDFP_a_00069

Appendix (Additional Tables)

Table A1: Cohort-Specific Average Treatment Effects of Third-Grade Reading Retention Policy on 4th Grade District RLA Achievement: Cohort-Specific ATET Estimates (AIPW)

Cohort	**ATET**	**Robust std. err.**	**z**	**P-value**
2015 Cohort				
year				
2011	0.008	0.028	0.300	0.766
2012	0.031	0.013	2.410	0.016
2013	0.022	0.011	2.060	0.040
2014	0.009	0.014	0.630	0.529
2015	0.017	0.037	0.450	0.653
2016	0.061	0.021	2.900	0.004
2017	0.058	0.025	2.340	0.019
2018	0.075	0.018	4.110	0.000
2019	0.081	0.023	3.490	0.000
2016 Cohort				
year				
2011	0.008	0.030	0.250	0.804
2012	0.005	0.010	0.540	0.589
2013	-0.002	0.023	-0.090	0.927
2014	0.082	0.013	6.210	0.000
2015	0.060	0.021	2.800	0.005
2016	0.024	0.020	1.190	0.234
2017	0.071	0.029	2.460	0.014
2018	0.108	0.027	3.940	0.000
2019	0.193	0.028	6.800	0.000

Note: ATET computed using covariates.

Note: The base time for pretreatment ATETs refers to the previous period.

Table A2. Evolution of Policy Effects on RLA Achievement: Dynamic ATETs by duration of exposure

Exposure	ATET	std. err.	z	P>z
-5	0.008	0.030	0.250	0.804
-4	0.008	0.023	0.340	0.733
-3	0.027	0.013	2.000	0.046
-2	0.029	0.012	2.410	0.016
-1	0.025	0.013	1.870	0.061
0	0.019	0.026	0.730	0.463
1	0.064	0.019	3.430	0.001
2	0.076	0.024	3.120	0.002
3	0.123	0.040	3.090	0.002
4	0.081	0.023	3.490	0.000

Note: The base time for pretreatment ATETs refers to the previous period.

Note: Exposure is the number of periods since the first treatment time.

Table A3. Dynamic Average Treatment Effects on the Treated (ATETs) by Cohort and Year

Cohort	ATET	Robust std. err.	z	P>z
2013				
year				
2011	0.004	0.023	0.180	0.854
2012	-0.018	0.020	-0.930	0.352
2013	-0.005	0.015	-0.340	0.736
2014	0.049	0.019	2.630	0.009
2015	0.136	0.035	3.930	0.000
2016	0.154	0.025	6.220	0.000
2017	0.205	0.029	7.150	0.000
2018	0.237	0.028	8.430	0.000
2019	0.322	0.036	8.930	0.000
2015				
year				
2011	0.010	0.030	0.320	0.746
2012	0.032	0.013	2.410	0.016
2013	0.022	0.012	1.770	0.076
2014	0.022	0.010	2.290	0.022
2015	0.024	0.038	0.620	0.533
2016	0.061	0.021	2.900	0.004
2017	0.058	0.025	2.340	0.019
2018	0.075	0.018	4.110	0.000
2019	0.081	0.023	3.490	0.000

Note: ATET computed using covariates.

Note: The base time for pretreatment ATETs refers to the previous period.

Table A4. Average Treatment Effects by Duration of Exposure to the Third-Grade Retention Policy (AIPW Estimates)

Exposure	ATET	Robust std. err.	z	P>z
-4	0.010	0.030	0.320	0.746
-3	0.032	0.013	2.410	0.016
-2	0.020	0.013	1.510	0.131
-1	0.008	0.015	0.500	0.614
0	0.014	0.026	0.520	0.601
1	0.057	0.016	3.470	0.001
2	0.086	0.033	2.630	0.009
3	0.107	0.030	3.530	0.000
4	0.125	0.041	3.070	0.002
5	0.237	0.028	8.430	0.000
6	0.322	0.036	8.930	0.000

Note: The base time for pretreatment ATETs refers to the previous period.

Note: Exposure is the number of periods since the first treatment time.

Table A5 Average Treatment Effects on the Treated (ATETs) of Third-Grade Retention Policy by Cohort and Year (AIPW Estimates)

Cohort	**ATET**	**Robust std. err**	**z**	**P> z**
2015				
year				
2011	0.013	0.017	0.760	0.446
2012	0.012	0.016	0.720	0.474
2013	-0.014	0.013	-1.040	0.299
2014	0.007	0.019	0.350	0.725
2015	0.006	0.036	0.170	0.862
2016	-0.020	0.043	-0.470	0.639
2017	-0.031	0.051	-0.600	0.546
2018	0.020	0.014	1.440	0.151
2019	0.033	0.028	1.150	0.250
2016				
year				
2011	0.007	0.016	0.430	0.669
2012	-0.031	0.007	-4.340	0.000
2013	-0.009	0.028	-0.310	0.753
2014	0.097	0.021	4.650	0.000
2015	0.049	0.038	1.290	0.198
2016	-0.030	0.022	-1.360	0.174
2017	-0.007	0.035	-0.210	0.835
2018	0.069	0.021	3.290	0.001
2019	0.129	0.003	37.350	0.000

Note: ATET computed using covariates.

Note: The base time for pretreatment ATETs refers to the previous period.

Table A6 Average Treatment Effects on the Treated (ATETs) by Duration of Exposure

Exposure	ATET	Robust std. err.	z	P>z
-5	0.007	0.016	0.430	0.669
-4	0.007	0.015	0.440	0.658
-3	0.009	0.016	0.550	0.581
-2	0.000	0.019	0.010	0.994
-1	0.020	0.019	1.060	0.290
0	-0.006	0.025	-0.230	0.819
1	-0.015	0.038	-0.400	0.686
2	0.005	0.041	0.120	0.903
3	0.064	0.034	1.900	0.057
4	0.033	0.028	1.150	0.250

Note: The base time for pretreatment ATETs refers to the previous period.

Note: Exposure is the number of periods since the first treatment time.